**用户申请评分卡开发**

成员：陈广威、梁锦伦

## 项目背景：

评分卡模型又叫做信用评分卡模型，最早由美国信用评分巨头FICO公司于20世纪60年代推出，在信用风险评估以及金融风险控制领域中广泛使用。银行利用评分卡模型对客户的信用历史数据的多个特征进行打分，得到不同等级的信用评分，从而判断客户的优质程度，据此决定是否准予授信以及授信的额度和利率。相较资深从业人员依靠自身的经验设置的专家规则，评分卡模型的使用具有很明显的优点：

1. 判断快速：系统只需要按照评分卡逐项打分，最后通过相应的公式计算出总分，即可准确判断出是否为客户授信以及额度和利率。
2. 客观透明：评分卡模型的标准是统一的，无论是客户还是风险审核人员，都可以通过评分卡一眼看出评分结果和评判依据。
3. 应用范围广：由于评分卡的评分项是客观计算，其得出的分数具有广泛的参考性和适用性。例如，生活中常见的支付宝芝麻信用分，就是依据评分卡模型计算得出。

评分卡模型在银行不同的业务阶段体现的方式和功能也不一样。按照借贷用户的借贷时间，评分卡模型可以划分为以下三种：

1. 贷前：申请评分卡（Application score card），又称为A卡。使用申请评分卡可以更准确地评估申请人的未来表现(违约率)，降低坏账率，加快(自动化)审批流程, 降低营运成本，还能增加审批决策的客观性和一致性，提高客户满意度
2. 贷中：行为评分卡（Behavior score card），又称为B卡。使用行为评分卡可以建立更好的客户管理策略, 提高赢利，减少好客户的流失，还能对可能拖欠的客户做出预警。
3. 贷后：催收评分卡（Collection score card），又称为C卡。使用催收评分卡可以优化催收策略，提高欠账的回收率，减少不必要的催收行为，降低营运成本。

## 项目流程：

信用风险评级模型的主要开发流程如下：

（1） 数据获取，包括获取存量客户及潜在客户的数据。存量客户是指已经在证券公司开展相关融资类业务的客户，包括个人客户和机构客户；潜在客户是指未来拟在证券公司开展相关融资类业务的客户，主要包括机构客户，这也是解决证券业样本较少的常用方法，这些潜在机构客户包括上市公司、公开发行债券的发债主体、新三板上市公司、区域股权交易中心挂牌公司、非标融资机构等。

（2） 数据预处理，主要工作包括数据清洗、缺失值处理、异常值处理，主要是为了将获取的原始数据转化为可用作模型开发的格式化数据。

（3） 探索性数据分析，该步骤主要是获取样本总体的大概情况，描述样本总体情况的指标主要有直方图、箱形图等。

（4） 变量选择，该步骤主要是通过统计学的方法，筛选出对违约状态影响最显著的指标。主要有单变量特征选择方法和基于机器学习模型的方法 。

（5） 模型开发，该步骤主要包括变量分段、变量的WOE（证据权重）变换和逻辑回归估算三部分。

（6） 模型评估，该步骤主要是评估模型的区分能力、预测能力、稳定性，并形成模型评估报告，得出模型是否可以使用的结论。

（7） 信用评分，根据逻辑回归的系数和WOE等确定信用评分的方法。将Logistic模型转换为标准评分的形式。

（8） 建立评分系统，根据信用评分方法，建立自动信用评分系统。



## 数据获取：

面向的客户群体：

（1） 存量客户数据（已开展相关融资类业务的客户，包括个人客户和机构客户，主要为个人客户）。

（2） 潜在客户数据（未来拟开展相关融资类业务的客户）。

数据集描述：

（1） 名称：application\_data.csv

（2） 来源：Kaggle上一个经典的评分卡案例（Loan Defaulter）

（3） 网址：<https://www.kaggle.com/gauravduttakiit/loan-defaulter>

（4） 内容：

基本属性：包括了借款人性别、职业、贷款类型等属性；

家庭情况：包括了借款人拥有的孩子数、家庭成员数等属性；

偿债能力：包括了借款人的收入、是否拥有房子、公寓、车等财产等属性；

联系方式：借款人是否提供电话、邮箱等属性；

信用往来：30天逾期次数、60天逾期次数等属性；

财产状况：包括客户居住建筑的标准化信息等属性；

贷款属性：包括客户是否提供某文件等属性；

其他因素：包括客户申请前一小时、一天等时间内向征信局查询客户的次数。

样本数量：307511条

特征数量：122个

详细的特征描述数据在文件columns\_description.csv中。

## 数据预处理：

大数据时代，要进行分析的数据量迅速膨胀，同时由于各种原因数据通常是不完整的（缺少某些感兴趣的属性值）、不一致的（包含代码或者名称的差异）、极易受到噪声（错误或异常值）的侵扰的。因为数据库太大，而且数据集经常来自多个异种数据源，低质量的数据将导致低质量的分析结果，所以分析数据所涉及的数据对象必须进行预处理。

（1） 目的：为了将获取的原始数据转化为可用作模型开发的格式化数据。

（2） 过程：特征清洗、缺失值处理、异常值处理

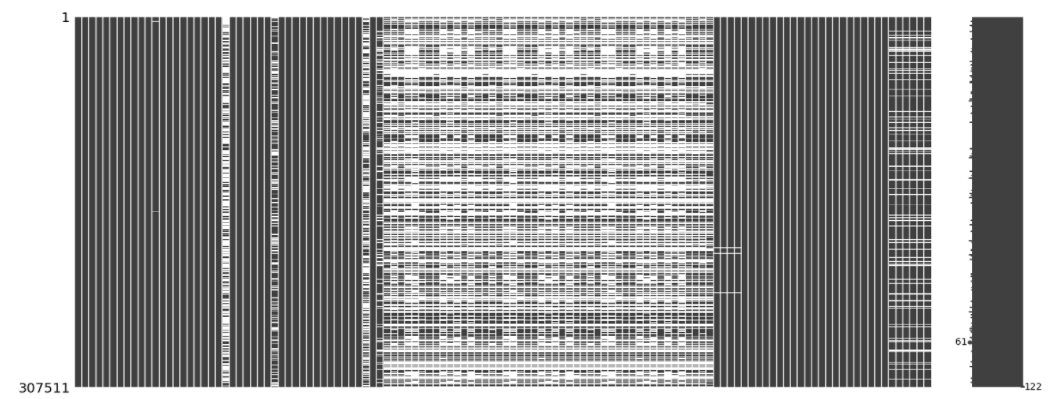
（A） 特征清洗：特征工程Dropping

（a） 统计特征的样本缺失率，若缺失值样本占总数比例极高（阈值），直接舍弃该特征，因为作为特征加入数据反而会引入噪声值。

技术：缺失值可视化（missingno）、无效矩阵的数据密集显示（matrix）、无效数据简单可视化的条形图（bar）、缺失性相关性热图（heatmap）、树状图（dendrogram）。

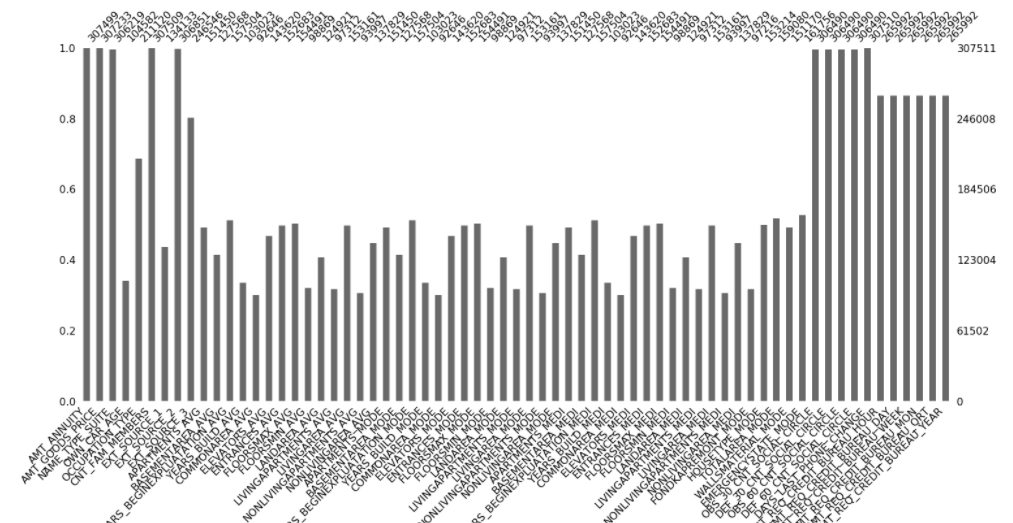
举例：

整体数据缺失值的无效矩阵数据密集显示（白色代表缺失）：



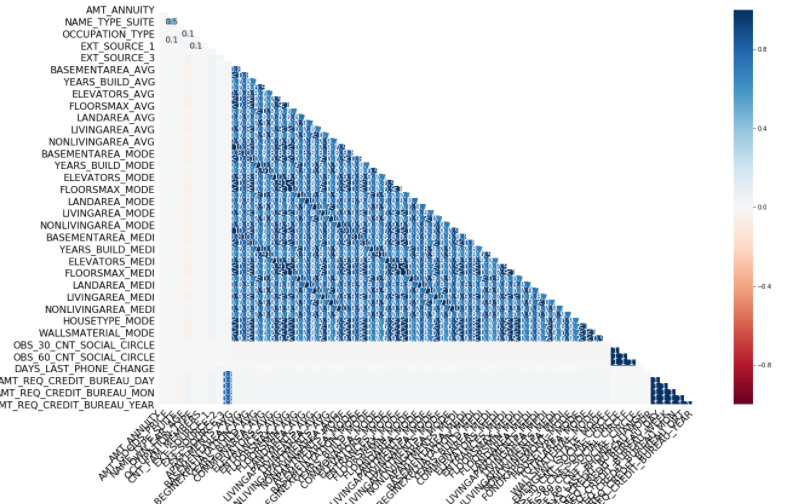
无效矩阵的长度分别代表122个特征变量，宽度代表307511个样本观测值，白色代表样本数据确实。从整体数据的无效矩阵数据显示图可以观察到，122个特征维度中存在较多特征显示为白线条，说明较多特征存在缺失值；其中不乏一些特征的白线条长度较长，说明一些特征样本数据缺失严重。

特征无效数据的条形图：



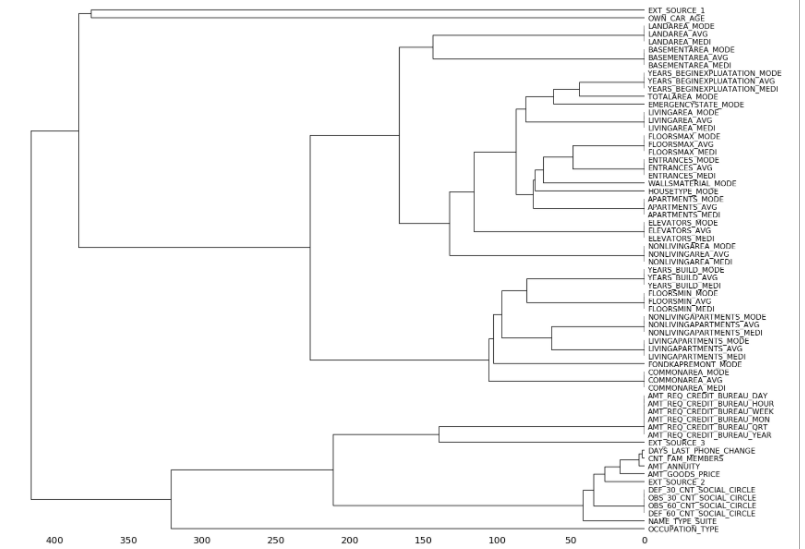
数据完整性条形图反映了存在缺失值的特征的非缺失样本的数量。从缺失特征的该条形图发现，存在缺失的特征中不少都是完整性低于0.5的，大部分特征完整性都小于0.6。

缺失性相关性热图：



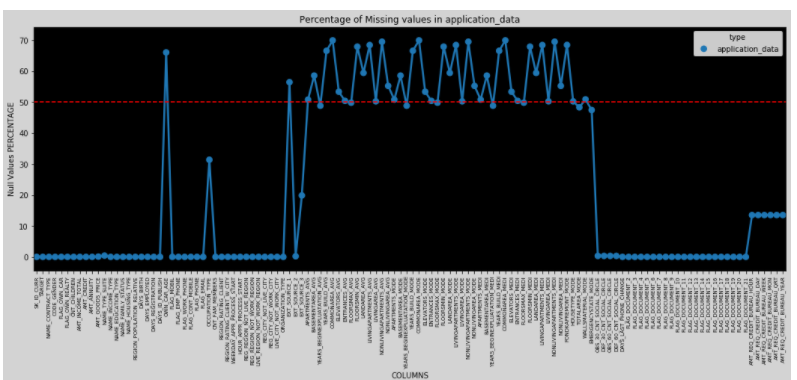
缺失性相关性热图可以观察出缺失特征之间数据缺失的相关性，相关性取值的绝对值越大，说明两特征之间数据的缺失越相关（颜色越深：蓝/红，缺失相关性越强）。若相关性为1，则一个特征数据的缺失必然会引起另一个特征数据的缺失；若相关性为-1，则一个特征数据的缺失必然不会引起另一个特征数据的缺失。观察上图可以发现，大部分缺失特征之间具有高正向相关性，右下角缺失特征之间的相关甚至接近1，说明特征的缺失具有彼此关联。根据缺失特征之间的缺失关联性，可以通过特征之间的关联对缺失值进行填充。

树状图：



上述树状图的横坐标即为缺失特征之间的距离，该距离表示特征的缺失率，越小说明缺失率越小。树状图会根据距离首先将缺失率最接近的两个特征进行聚类，而后逐步聚类，直至所有的特征聚为一类。不仅反映出特征的缺失率高低，同时反映了特征之间缺失率的相似度。

删除特征变量样本缺失率高于40%的特征：



红色横虚线是50%的缺失率阈值，整体数据中存在大量特征的数据缺失率高于50%。且还有很多特征的数据缺失率在48%和49%左右，这些特征的缺失值之大，对训练模型作用很小，甚至造成不利的作用，在考察了这些特征的性质后，对样本缺失率高达40%的数据进行了删除。

删除的特征名单为：

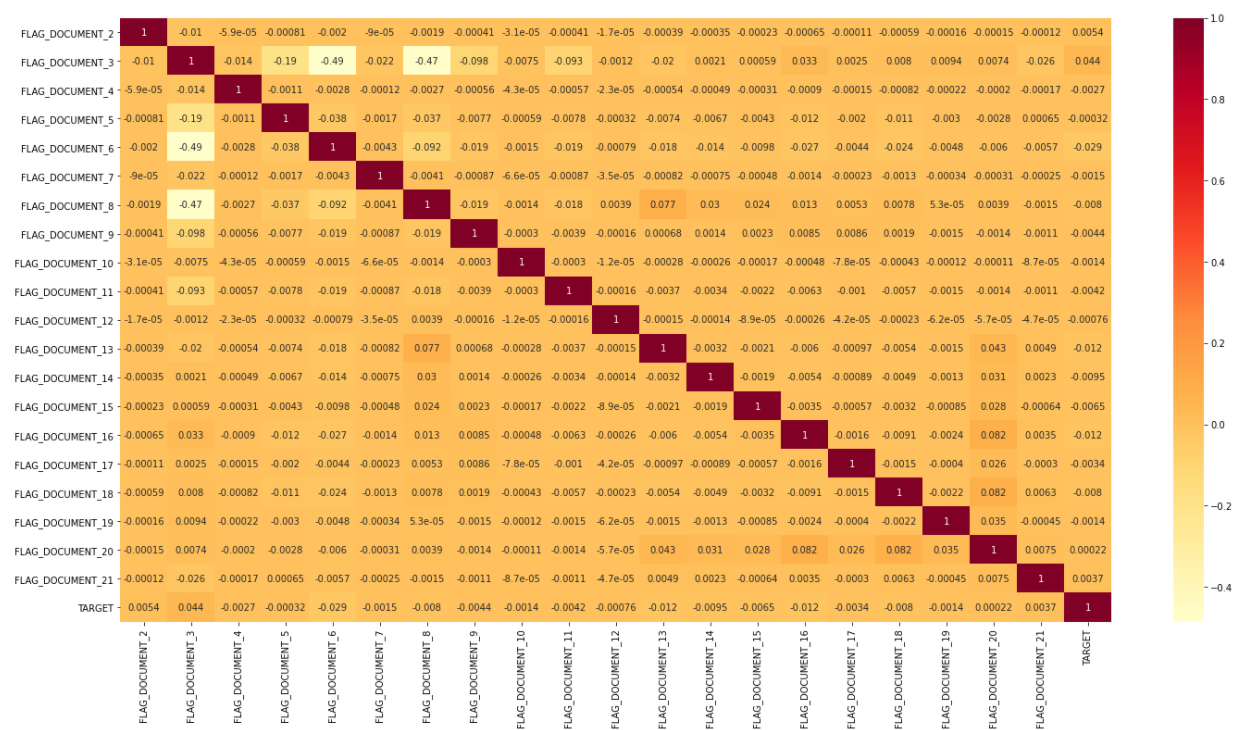
'OWN\_CAR\_AGE','EXT\_SOURCE\_1','APARTMENTS\_AVG','BASEMENTAREA\_AVG','YEARS\_BEGINEXPLUATATION\_AVG','YEARS\_BUILD\_AVG','COMMONAREA\_AVG','ELEVATORS\_AVG','ENTRANCES\_AVG','FLOORSMAX\_AVG','FLOORSMIN\_AVG','LANDAREA\_AVG','LIVINGAPARTMENTS\_AVG','LIVINGAREA\_AVG','NONLIVINGAPARTMENTS\_AVG','NONLIVINGAREA\_AVG','APARTMENTS\_MODE','BASEMENTAREA\_MODE','YEARS\_BEGINEXPLUATATION\_MODE','YEARS\_BUILD\_MODE','COMMONAREA\_MODE','ELEVATORS\_MODE','ENTRANCES\_MODE','FLOORSMAX\_MODE','FLOORSMIN\_MODE','LANDAREA\_MODE','LIVINGAPARTMENTS\_MODE','LIVINGAREA\_MODE','NONLIVINGAPARTMENTS\_MODE','NONLIVINGAREA\_MODE','APARTMENTS\_MEDI','BASEMENTAREA\_MEDI','YEARS\_BEGINEXPLUATATION\_MEDI','YEARS\_BUILD\_MEDI','COMMONAREA\_MEDI','ELEVATORS\_MEDI','ENTRANCES\_MEDI', 'FLOORSMAX\_MEDI', 'FLOORSMIN\_MEDI', 'LANDAREA\_MEDI', 'LIVINGAPARTMENTS\_MEDI','LIVINGAREA\_MEDI','NONLIVINGAPARTMENTS\_MEDI','NONLIVINGAREA\_MEDI','FONDKAPREMONT\_MODE','HOUSETYPE\_MODE','TOTALAREA\_MODE','WALLSMATERIAL\_MODE','EMERGENCYSTATE\_MODE'

（b） 通过目标变量与自变量特征之间的相关性，结合数据的分箱统计和逻辑分析，对无关特征进行删除。

技术：数据相关性热图（heatmap）、分箱统计条形图（countplot）。

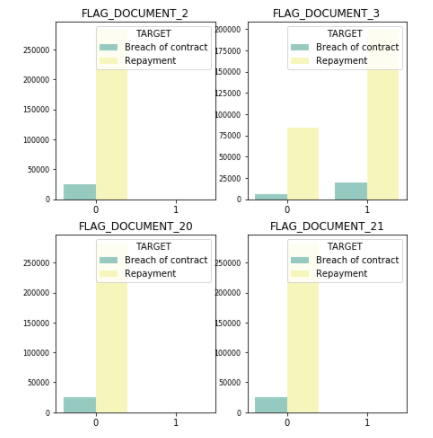
举例：

与提供相关文件有关的特征"FLAG\_DOCUMENT\_2","FLAG\_DOCUMENT\_3","FLAG\_DOCUMENT\_4","FLAG\_DOCUMENT\_5","FLAG\_DOCUMENT\_6","FLAG\_DOCUMENT\_7","FLAG\_DOCUMENT\_8","FLAG\_DOCUMENT\_9","FLAG\_DOCUMENT\_10","FLAG\_DOCUMENT\_11","FLAG\_DOCUMENT\_12","FLAG\_DOCUMENT\_13","FLAG\_DOCUMENT\_14","FLAG\_DOCUMENT\_15","FLAG\_DOCUMENT\_16","FLAG\_DOCUMENT\_17","FLAG\_DOCUMENT\_18","FLAG\_DOCUMENT\_19","FLAG\_DOCUMENT\_20","FLAG\_DOCUMENT\_21"与TARGET的相关性热图：



相关性热图反映了变量彼此之间的相关性。值为正则为正相关，为负则为负相关，绝对值越大，则相应的相关性越强。颜色越深，正相关性越强。热图显示，这些特征大部分与目标特征TARGET相关性很弱，对模型预测结果影响不大，为了简化工作，对除了FLAG\_DOCUMENT\_2、FLAG\_DOCUMENT\_3、FLAG\_DOCUMENT\_20、FLAG\_DOCUMENT\_21以外的与额外提供文件有关的特征进行删除。

FLAG\_DOCUMENT\_2、FLAG\_DOCUMENT\_3、FLAG\_DOCUMENT\_20、FLAG\_DOCUMENT\_21与TARGET的分箱统计条形图：



分箱统计条形图用于统计分类型变量不同取值时目标变量的取值情况，可分析分类型变量的取值对目标变量的影响。FLAG\_DOCUMENT\_2、FLAG\_DOCUMENT\_20、FLAG\_DOCUMENT\_21变量下，用户的违约与还款均发生在变量取值为0时，分析FLAG\_DOCUMENT\_2、FLAG\_DOCUMENT\_20、FLAG\_DOCUMENT\_21变量无法得出其取值为1时对目标变量的影响。所以我们对特征FLAG\_DOCUMENT\_2、FLAG\_DOCUMENT\_20、FLAG\_DOCUMENT\_21进行了删除。

我们也对'FLAG\_MOBIL', 'FLAG\_EMP\_PHONE', 'FLAG\_WORK\_PHONE', 'FLAG\_CONT\_MOBILE', 'FLAG\_PHONE', 'FLAG\_EMAIL'进行了相关性热图的分析以及分箱统计条形图的分析，发现这些变量对目标变量的相关性不大，同时主观上认为这些个人提供的联系方式与违约概率关系不大，因为一般来说，能成功申请到贷款的个人联系方式应该一般为真，故对这些特征进行了删除。

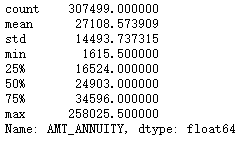
（B） 缺失值处理：特征工程Imputation、An extension to imputation

（a） 连续性特征、二值特征根据同特征样本之间的相似性，利用平均数、众数、中位数填充填补缺失值。

技术：描述性统计（describe）、箱线图（boxplot）

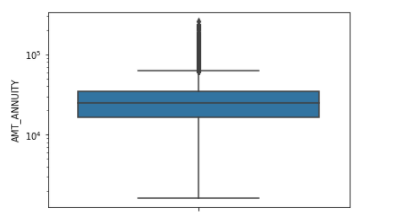
举例：

AMT\_ANNUITY的描述性统计：



描述性统计反映了数值型特征的计数、均值、标准差、最小值、25%分位数、50%分位数即中位数、75%分位数和最大值。

AMT\_ANNUITY的箱线图：



箱线图最下方的横线为该特征的下边缘，上方横线为其上边缘，中间蓝色矩形的上下边分别为其上四分位数和下四分位数，矩形中间的线为中位数，上下边缘之外的点为异常值。AMT\_ANNUITY存在较多异常值超出上限，因此使用中位数填充缺失值。

同时AMT\_GOODS\_PRICE、EXT\_SOURCE\_3、EXT\_SOURCE\_2、CNT\_FAM\_MEMBERS等特征经过分析同样通过中位数填充。

对OBS\_30\_CNT\_SOCIAL\_CIRCLE、OBS\_60\_CNT\_SOCIAL\_CIRCLE、AMT\_REQ\_CREDIT\_BUREAU\_YEAR这些中位数过大的特征选择了平均数填充。

对DEF\_30\_CNT\_SOCIAL\_CIRCLE、DEF\_60\_CNT\_SOCIAL\_CIRCLE、DAYS\_LAST\_PHONE\_CHANGE、DAYS\_LAST\_PHONE\_CHANGE、AMT\_REQ\_CREDIT\_BUREAU\_HOUR、AMT\_REQ\_CREDIT\_BUREAU\_DAY、AMT\_REQ\_CREDIT\_BUREAU\_MON、AMT\_REQ\_CREDIT\_BUREAU\_QRT 这些数据较分散的特征选择众数进行填充。

（b） 非连续性特征（分类值）将属性值进行分类，作为新的特征加入到数据集。

OCCUPATION\_TYPE、NAME\_TYPE\_SUITE等特征将缺失值作为一种属性填充缺失值。

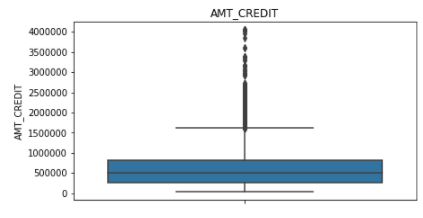
（C） 异常值处理

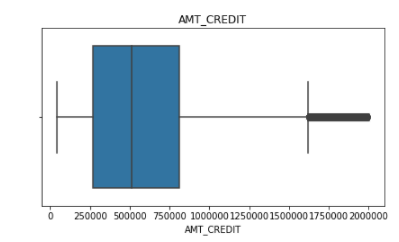
对连续性特征由于人为错误、数据处理错误等原因产生的异常值进行删除处理。

技术：箱线图（boxplot）、直方图（distplot）、核密度估计（kde）

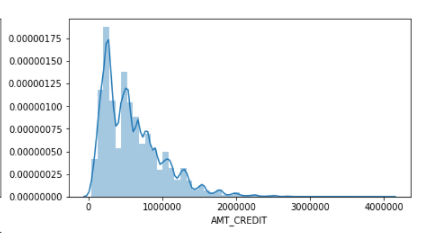
举例：

AMT\_CREDIT的箱线图：





AMT\_CREDIT的直方图和核密度估计：

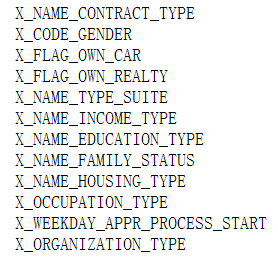


根据AMT\_CREDIT的箱线图和核密度曲线，可以发现绝大多数数据位于2000000以内，可以对超过2000000的数据进行删除。

同时，通过分析对AMT\_INCOME\_TOTAL超过800000的数据进行删除，对AMT\_ANNUITY超过100000的数据进行删除，对AMT\_GOODS\_PRICE超过2000000的数据进行删除，对CODE\_GENDER属性为XNA的数据进行删除。

## 探索性分析：

1. 特征转换:为了方便表示和进一步的处理，将非数值型特征进行编码，转换成数字，方便识别和后续的哑变量变换。生成以X\_前缀开头的变量。



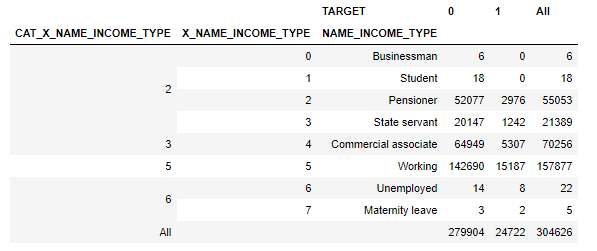
1. 数据分箱：局部平滑方法，通过考察“邻居”（周围的值）来平滑存储数据的值，即连续变量离散化，达到去噪，增加粒度，提高变量的可解释性和鲁棒性的目的。

方法：决策树最优分箱（使用sklearn提供的决策树，将内部节点的阈值作为分箱的切点，利用变量值x和目标变量y之间产生一个映射关系，从而将特征用树模型拟合目标变量）

意义：数据分箱要求变量分箱后每箱的坏样本占比具有一定的区分度。决策树分箱通过信息增益或者gini系数进行分裂特征，信息增益或者gini系数也可以说成分裂点最大可能的提升分裂前后变量的区分度，所以分箱的每个分裂点都是根据目标变量选择的最优分裂方式。

举例：

列联表显示部分分箱结果：



NAME\_INCOME\_TYPE为原始特征名称，X\_NAME\_INCOME\_TYPE为独热编码后的衍生的特征，CAT\_X\_NAME\_INCOME\_TYPE为分箱后的结果。数据分箱要求每个箱子的坏样本具有明显的区分度，上述分箱结果（2箱子：0，1，2，3；3箱子：4；5箱子：5；6箱子：6，7）通过简单计算就可以得出，每个箱子的坏样本具有明显的区分度。

使用决策树将变量进行分箱，最后我们得到了CAT\_开头的新特征。

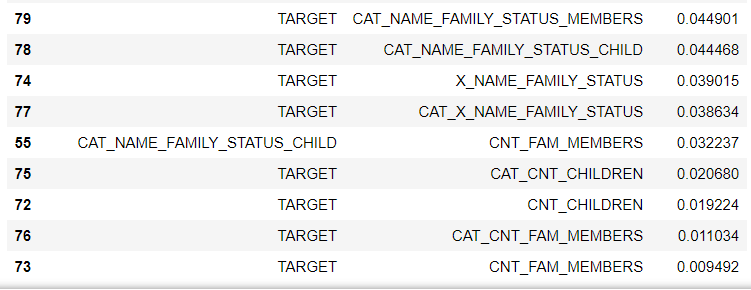
1. 特征分析和特征衍生

针对特征数量大的问题，我们将保留的特征分成家庭相关、住址相关、收入和工作相关、联系方式相关、贷款相关和其他等类别来处理，使用的方法包括对一类特征变量（如工作/收入相关的特征变量）及其分箱结果与目标变量TARGET之间进行描述性统计，从而初步分析变量与目标变量之间是否存在相关性影响。根据特征分析结果，采用技术衍生的手段利用现有的特征进行某种组合衍生成新的特征，再结合特征分析判断衍生的特征对目标变量是否存在较强的影响。

1. 家庭相关

与家庭相关的特征变量包括:CNT\_CHILDREN、NAME\_TYPE\_SUITE、NAME\_FAMILY\_STATUS、CNT\_FAM\_MEMBERS、X\_NAME\_TYPE\_SUITE、X\_NAME\_FAMILY\_STATUS、CAT\_CNT\_CHILDREN、CAT\_CNT\_FAM\_MEMBERS、CAT\_X\_NAME\_TYPE\_SUITE、CAT\_X\_NAME\_FAMILY\_STATUS。

为了从特征中获得更多信息，我们通过决策树结合CAT\_X\_NAME\_FAMILY\_STATUS和CNT\_CHILDREN生成一个包含家庭状况和孩子数量信息的新特征CAT\_NAME\_FAMILY\_STATUS\_CHILD，结合NAME\_FAMILY\_STATUS、CNT\_FAM\_MEMBERS、CNT\_CHILDREN来生成一个包含家庭状况、家庭成员数量、孩子数量的特征CAT\_NAME\_FAMILY\_STATUS\_MEMBERS。

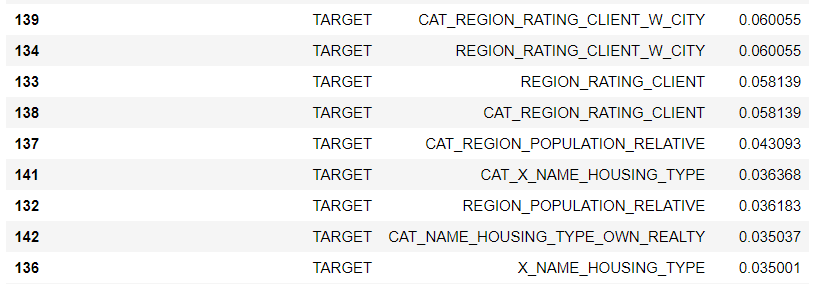


通过相关性分析，我们发现CAT\_NAME\_FAMILY\_STATUS\_CHILD和CAT\_NAME\_FAMILY\_STATUS\_MEMBERS的相关性高达0.995613，而两者与TARGET相关性相差不大，但在所有家庭相关的特征中相关性是最大的，所以最后我们选择保留 CAT\_NAME\_FAMILY\_STATUS\_CHILD。

1. 住址相关

与住址相关的特征包括： FLAG\_OWN\_REALTY、 NAME\_HOUSING\_TYPE、REGION\_POPULATION\_RELATIVE、REGION\_RATING\_CLIENT、 REGION\_RATING\_CLIENT\_W\_CITY、X\_FLAG\_OWN\_REALTY、X\_NAME\_HOUSING\_TYPE、CAT\_REGION\_POPULATION\_RELATIVE、CAT\_REGION\_RATING\_CLIENT、 CAT\_REGION\_RATING\_CLIENT\_W\_CITY、CAT\_X\_FLAG\_OWN\_REALTY、 CAT\_X\_NAME\_HOUSING\_TYPE。

为了从特征中获得更多的信息，我们通过决策树结合CAT\_X\_NAME\_HOUSING\_TYPE和CAT\_X\_FLAG\_OWN\_REALTY生成一个包含住房情况信息和是否拥有房子或公寓信息的特征 CAT\_NAME\_HOUSING\_TYPE\_OWN\_REALTY。

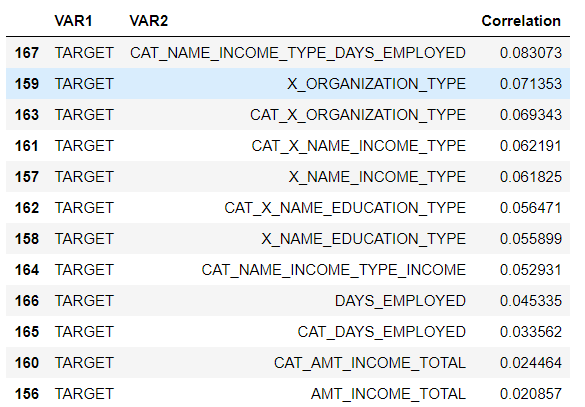


接着通过相关性分析，我们发现目标变量TARGET与CAT\_REGION\_RATING\_CLIENT\_W\_CITY相关性较大，之后是CAT\_REGION\_RATING\_CLIENT、CAT\_REGION\_POPULATION\_RELATIVE（负相关），而CAT\_REGION\_RATING\_CLIENT\_W\_CITY与CAT\_REGION\_RATING\_CLIENT、CAT\_REGION\_POPULATION\_RELATIVE高度相关，CAT\_REGION\_RATING\_CLIENT、CAT\_REGION\_POPULATION\_RELATIVE互相关联，因此保留CAT\_REGION\_RATING\_CLIENT\_W\_CITY即可。而TARGET与CAT\_X\_NAME\_HOUSING\_TYPE相关性较大，CAT\_X\_NAME\_HOUSING\_TYPE和CAT\_REGION\_RATING\_CLIENT\_W\_CITY的相关性较低，所以保留CAT\_X\_NAME\_HOUSING\_TYPE。最终保留CAT\_REGION\_RATING\_CLIENT\_W\_CITY、CAT\_X\_NAME\_HOUSING\_TYPE

1. 收入/工作相关

与收入/工作相关的特征包括：AMT\_INCOME\_TOTAL、NAME\_INCOME\_TYPE、NAME\_EDUCATION\_TYPE、X\_NAME\_INCOME\_TYPE、X\_NAME\_EDUCATION\_TYPE、X\_ORGANIZATION\_TYPE、CAT\_AMT\_INCOME\_TOTAL、CAT\_X\_NAME\_INCOME\_TYPE、CAT\_X\_NAME\_EDUCATION\_TYPE、CAT\_X\_ORGANIZATION\_TYPE、CAT\_DAYS\_EMPLOYED、DAYS\_EMPLOYED。

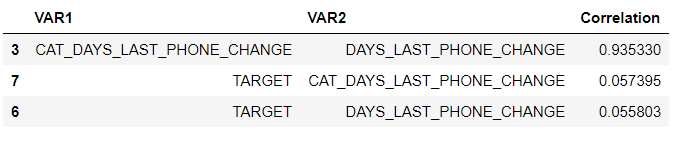
为了从特征中获得更多的信息，我们通过决策树结合CAT\_X\_NAME\_INCOME\_TYPE和AMT\_INCOME\_TOTAL生成一个包含客户收入类型信息和客户收入信息的特征CAT\_NAME\_INCOME\_TYPE\_INCOME，通过结合CAT\_X\_NAME\_INCOME\_TYPE和DAYS\_EMPLOYED生成一个客户收入类型信息和在申请贷款前多久开始工作信息的新特征CAT\_NAME\_INCOME\_TYPE\_DAYS\_EMPLOYED。



通过相关性分析，我们发现TARGET与CAT\_NAME\_INCOME\_TYPE\_DAYS\_EMPLOYED、X\_ORGANIZATION\_TYPE、CAT\_X\_NAME\_INCOME\_TYPE、CAT\_X\_NAME\_EDUCATION\_TYPE、 CAT\_AMT\_INCOME\_TOTAL之间相关性较大，根据这些变量之间的相关性，保留CAT\_NAME\_INCOME\_TYPE\_DAYS\_EMPLOYED和CAT\_X\_NAME\_EDUCATION\_TYPE。

1. 联系方式相关

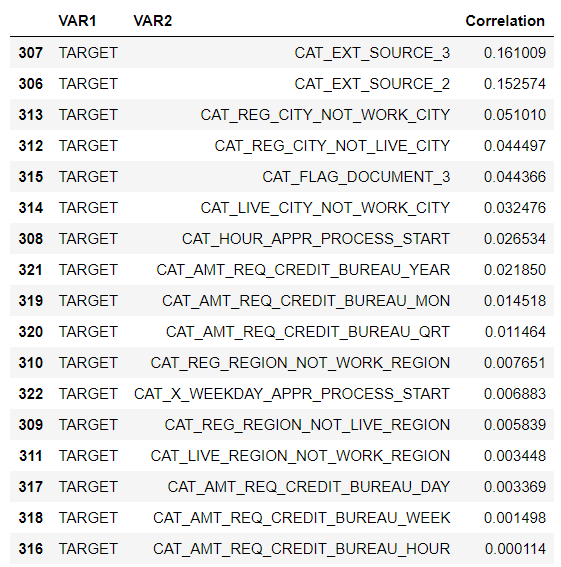
与联系方式相关的特征包括：DAYS\_LAST\_PHONE\_CHANGE、CAT\_DAYS\_LAST\_PHONE\_CHANGE。



通过相关性分析，我们发现TARGET与CAT\_DAYS\_LAST\_PHONE\_CHANGE、DAYS\_LAST\_PHONE\_CHANGE相关性较大，但同时CAT\_DAYS\_LAST\_PHONE\_CHANGE和DAYS\_LAST\_PHONE\_CHANGE之间具有高度相关性，因此保留CAT\_DAYS\_LAST\_PHONE\_CHANGE即可。

1. 贷款相关

与贷款相关的变量包括：WEEKDAY\_APPR\_PROCESS\_START、HOUR\_APPR\_PROCESS\_START、REG\_REGION\_NOT\_LIVE\_REGION、REG\_REGION\_NOT\_WORK\_REGION、LIVE\_REGION\_NOT\_WORK\_REGION、REG\_CITY\_NOT\_LIVE\_CITY、REG\_CITY\_NOT\_WORK\_CITY、LIVE\_CITY\_NOT\_WORK\_CITY、OBS\_30\_CNT\_SOCIAL\_CIRCLE、DEF\_30\_CNT\_SOCIAL\_CIRCLE、OBS\_60\_CNT\_SOCIAL\_CIRCLE、DEF\_60\_CNT\_SOCIAL\_CIRCLE、AMT\_REQ\_CREDIT\_BUREAU\_HOUR、AMT\_REQ\_CREDIT\_BUREAU\_DAY、AMT\_REQ\_CREDIT\_BUREAU\_WEEK、AMT\_REQ\_CREDIT\_BUREAU\_MON、AMT\_REQ\_CREDIT\_BUREAU\_QRT、AMT\_REQ\_CREDIT\_BUREAU\_YEAR、X\_WEEKDAY\_APPR\_PROCESS\_START、CAT\_HOUR\_APPR\_PROCESS\_START、CAT\_REG\_REGION\_NOT\_LIVE\_REGION、CAT\_REG\_REGION\_NOT\_WORK\_REGION、CAT\_LIVE\_REGION\_NOT\_WORK\_REGION、CAT\_REG\_CITY\_NOT\_LIVE\_CITY、CAT\_REG\_CITY\_NOT\_WORK\_CITY、CAT\_LIVE\_CITY\_NOT\_WORK\_CITY、CAT\_OBS\_30\_CNT\_SOCIAL\_CIRCLE、CAT\_DEF\_30\_CNT\_SOCIAL\_CIRCLE、CAT\_OBS\_60\_CNT\_SOCIAL\_CIRCLE、CAT\_DEF\_60\_CNT\_SOCIAL\_CIRCLE、CAT\_AMT\_REQ\_CREDIT\_BUREAU\_HOUR、CAT\_AMT\_REQ\_CREDIT\_BUREAU\_DAY、CAT\_AMT\_REQ\_CREDIT\_BUREAU\_WEEK、CAT\_AMT\_REQ\_CREDIT\_BUREAU\_MON、CAT\_AMT\_REQ\_CREDIT\_BUREAU\_QRT、CAT\_AMT\_REQ\_CREDIT\_BUREAU\_YEAR、CAT\_X\_WEEKDAY\_APPR\_PROCESS\_START、EXT\_SOURCE\_2、EXT\_SOURCE\_3、CAT\_EXT\_SOURCE\_2、CAT\_EXT\_SOURCE\_3。

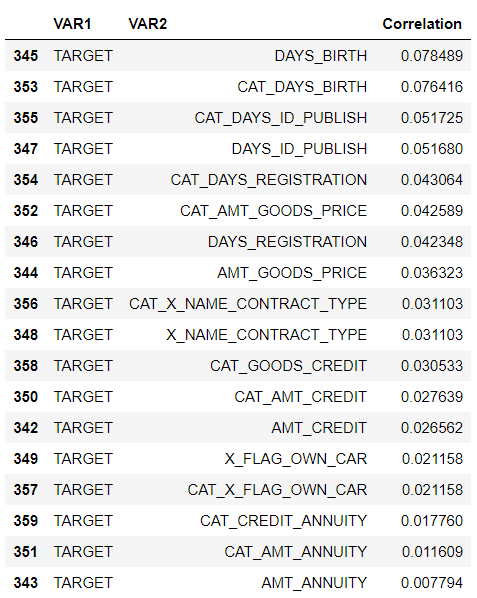


通过相关性分析，我们发现Target和CAT\_EXT\_SOURCE\_2, CAT\_EXT\_SOURCE\_3,CAT\_REG\_CITY\_NOT\_WORK\_CITY的相关性最高，其次是CAT\_REG\_CITY\_NOT\_LIVE\_CITY、CAT\_FLAG\_DOCUMENT\_3，之后是CAT\_LIVE\_CITY\_NOT\_WORK\_CITY。CAT\_REG\_CITY\_NOT\_WORK\_CITY和CAT\_FLAG\_DOCUMENT\_3的相关性很低，CAT\_REG\_CITY\_NOT\_WORK\_CITY和CAT\_REG\_CITY\_NOT\_LIVE\_CITY、CAT\_LIVE\_CITY\_NOT\_WORK\_CITY的相关性很高。最后我们选择保留CAT\_REG\_CITY\_NOT\_WORK\_CITY、CAT\_FLAG\_DOCUMENT\_3,CAT\_EXT\_SOURCE\_2, CAT\_EXT\_SOURCE\_3。

1. 其他特征

其他特征包括：NAME\_CONTRACT\_TYPE、FLAG\_OWN\_CAR、AMT\_CREDIT、AMT\_ANNUITY、AMT\_GOODS\_PRICE、DAYS\_BIRTH、DAYS\_REGISTRATION、DAYS\_ID\_PUBLISH、X\_NAME\_CONTRACT\_TYPE、X\_FLAG\_OWN\_CAR、CAT\_AMT\_CREDIT、CAT\_AMT\_ANNUITY、CAT\_AMT\_GOODS\_PRICE、CAT\_DAYS\_BIRTH、CAT\_DAYS\_REGISTRATION、CAT\_DAYS\_ID\_PUBLISH、CAT\_X\_NAME\_CONTRACT\_TYPE、CAT\_X\_FLAG\_OWN\_CAR。

为了从特征中获得更多的信息，我们通过决策树将AMT\_GOODS\_PRICE和AMT\_CREDIT生成一个包含贷款价格和贷款额信息的特征CAT\_GOODS\_CREDIT，结合AMT\_CREDIT和AMT\_ANNUITY生成一个包含贷款额和贷款年金信息的CAT\_CREDIT\_ANNUITY。



通过相关性分析，我们发现TARGET和DAYS\_BIRTH、CAT\_DAYS\_BIRTH、CAT\_DAYS\_ID\_PUBLISH、DAYS\_ID\_PUBLISH、CAT\_DAYS\_REGISTRATION、CAT\_AMT\_GOODS\_PRICE具有较高的相关性。最后我们选择保留：CAT\_DAYS\_BIRTH、CAT\_DAYS\_REGISTRATION、CAT\_DAYS\_ID\_PUBLISH、CAT\_AMT\_GOODS\_PRICE、CAT\_X\_NAME\_CONTRACT\_TYPE、CAT\_X\_FLAG\_OWN\_CAR。

## 变量选择：

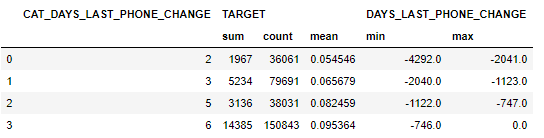
通过统计学的方法，筛选出对目标变量即违约与否影响最显著的指标。

（1） 特征分析

通过特征变量与目标变量之间的描述性统计，人为分析特征变量与目标变量之间的关联性，判断该特征是否对目标变量有影响，作为变量的初步筛选。

举例：

CAT\_DAYS\_LAST\_PHONE\_CHANGE与TARGET、DAYS\_LAST\_PHONE\_CHANGE之间的描述性统计：



最佳更换电话号码越频繁，最终违约概率越大，因此特征与目标变量之间相关。

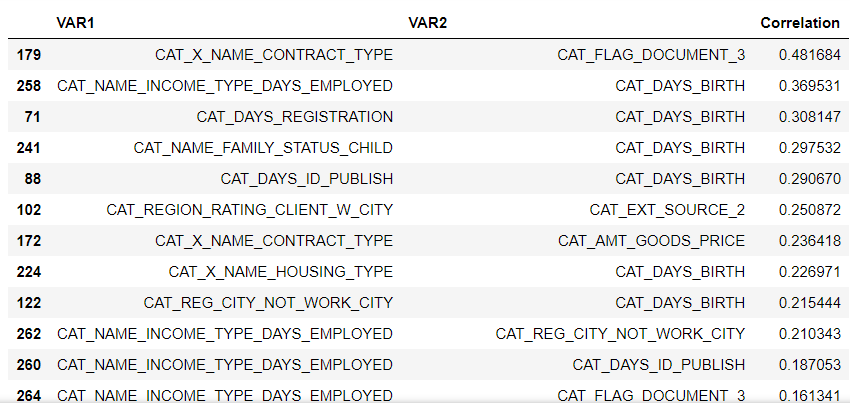
（2） 双样本比例Z检验

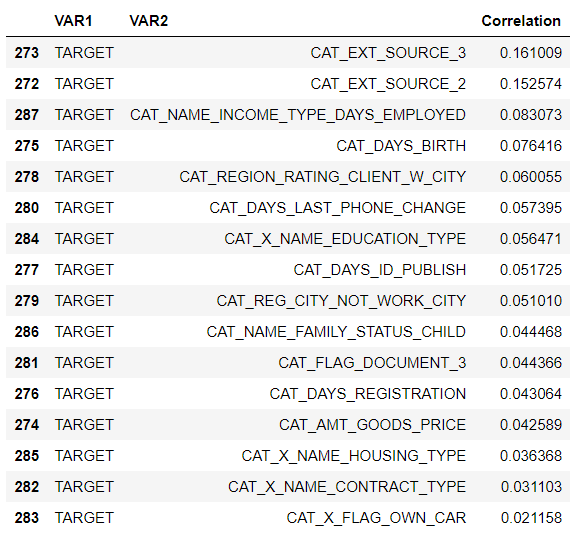
对特征分析中描述统计为两种情况的特征变量进行对目标变量的比例检验，分析该特征变量对目标变量是否存在显著影响。

技术：statsmodels.stats.proportion.proportions\_ztest

（3） 相关性分析

使用相关性矩阵得出特征变量之间、特征变量与目标变量之间的相关性，分析特征变量与目标变量之间的相关性，选择与目标变量相关性高的特征变量作为已选择特征；分析已选择特征变量之间的相关性，对其中彼此相关性高的两个或多个变量进行选择剔除，尽量确保已选择的变量之间具有低相关性，减少变量间的关联，避免变量重复、多重共线性问题和过拟合问题。





由图可知，CAT\_X\_NAME\_CONTRACT\_TYPE和CAT\_FLAG\_DOCUMENT\_3的相关性为0.481684，这意味着两个关联大，删除CAT\_FLAG\_DOCUMENT\_3； CAT\_DAYS\_BIRTH和CAT\_DAYS\_REGISTRATION、CAT\_NAME\_FAMILY\_STATUS\_CHILD、CAT\_DAYS\_ID\_PUBLISH相关性均为0.3左右，而CAT\_DAYS\_BIRTH和CAT\_NAME\_INCOME\_TYPE\_DAYS\_EMPLOYED的相关性为0.369，选择删除CAT\_DAYS\_BIRTH，减少变量间的关联。

（4） 变量的评价

（A） WOE（Weight of Evidence）：证据权重类似于信息熵反映了自变量取值对目标变量的影响。WOE越大表示该特征正向作用越大，它对数据进行了归一化处理，也就是将所有不同特征划在了统一的尺度上。公式：WOE=ln(good attribute/bad attribute)。

（B） IV（Information Value）：WOE没有考虑分组中样本占整体样本的比例，如果一个分组的WOE值很高，但是样本数占整体样本数很低，则对变量整体预测的能力会下降。IV信息量考虑了分组中样本占整体样本的比例，相当于WOE的加权求和。计算特征内部的信息含量，如果信息含量足够大，则表示其为有价值的特征。

（一般）标准：

小于0.02：预测能力极低（Unpredictive）

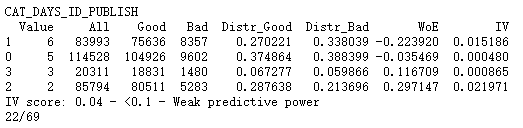
0.02到0.1之间：预测能力较低（Weak）

0.1到0.3之间：预测能力中等（Medium

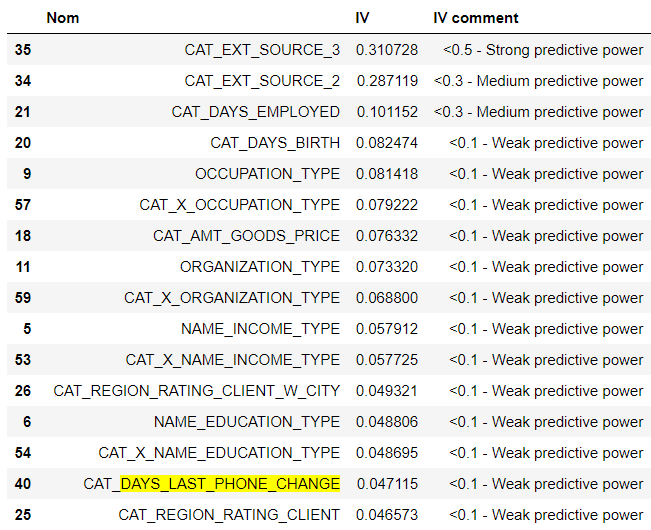
0.3到0.5之间：预测能力较强（Strong）

举例：

CAT\_DAYS\_ID\_PUBLISH的WOE和IV值：



上图展示了DAYS\_ID\_PUBLISH的WOE、IV情况，Value表示DAYS\_ID\_PUBLISH的分箱情况即CAT\_DAYS\_ID\_PUBLISH的取值；All表示该Value对应的计数；Good表示All计数中最终TARGET为0即不违约的计数；Bad则为All中最终违约的计数；Distr\_Good为该Value下不违约的数量占总不违约计数（Good总数）的比例；Distr\_Bad为Value下违约的数量占总违约计数（Bad总数）的比例；WOE则是，衡量了该Value组内违约客户、不违约客户的比值和所有样本中这个比值的差异；IV值为，衡量变量的预测能力。



通过IV值，我们发现CAT\_EXT\_SOURCE\_3和CAT\_EXT\_SOURCE\_2和CAT\_DAYS\_EMPLOYED提供的信息价值最大。

最后，我们选择的特征包括：'CAT\_EXT\_SOURCE\_2','CAT\_EXT\_SOURCE\_3','CAT\_AMT\_GOODS\_PRICE','CAT\_DAYS\_REGISTRATION','CAT\_DAYS\_ID\_PUBLISH','CAT\_REGION\_RATING\_CLIENT\_W\_CITY','CAT\_REG\_CITY\_NOT\_WORK\_CITY','CAT\_DAYS\_LAST\_PHONE\_CHANGE','CAT\_X\_NAME\_CONTRACT\_TYPE','CAT\_X\_FLAG\_OWN\_CAR','CAT\_X\_NAME\_EDUCATION\_TYPE','CAT\_X\_NAME\_HOUSING\_TYPE','CAT\_NAME\_FAMILY\_STATUS\_CHILD','CAT\_NAME\_INCOME\_TYPE\_DAYS\_EMPLOYED'。

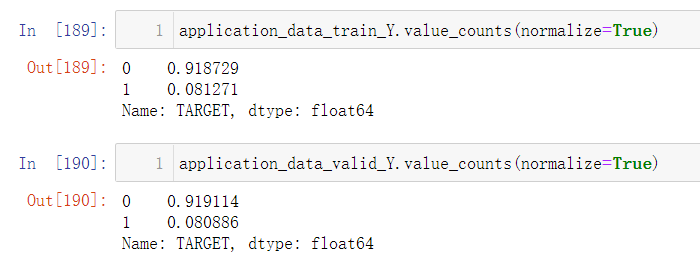
**模型开发：**

1. 变量离散化

我们将处理好的数据划分成自变量数据集和目标变量数据集，同时对自变量通过哑变量生产将变量离散化。

1. 划分数据集

通过sklearn中train\_test\_split方法，我们将70%的数据作为训练集，将30%的数据作为验证集。

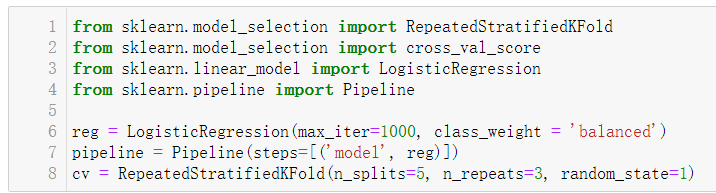


从上图可知，训练集和验证集中目标变量的违约与不违法的比例几乎一致，所以训练集和验证集的划分是很均匀的。

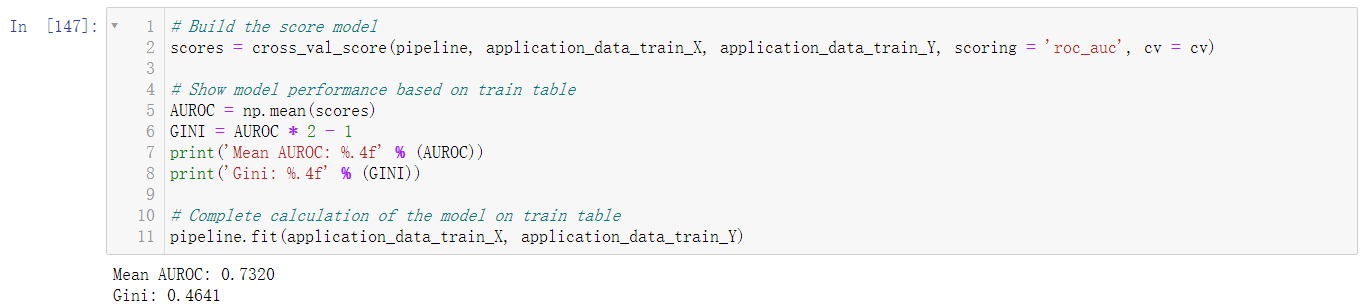
1. Logistic回归模型：

逻辑回归是一种用于解决二分类问题的机器学习方法，用于估计某种事物的可能性。这里的“可能性”不等同于数学上的“概率”，它的结果往往用与和其他特征值加权求和。所以运用逻辑回归模型，可以给我们选择的特征赋予权重，再对特征加权求和得到评分。

我们首先采用逻辑回归模型去建模



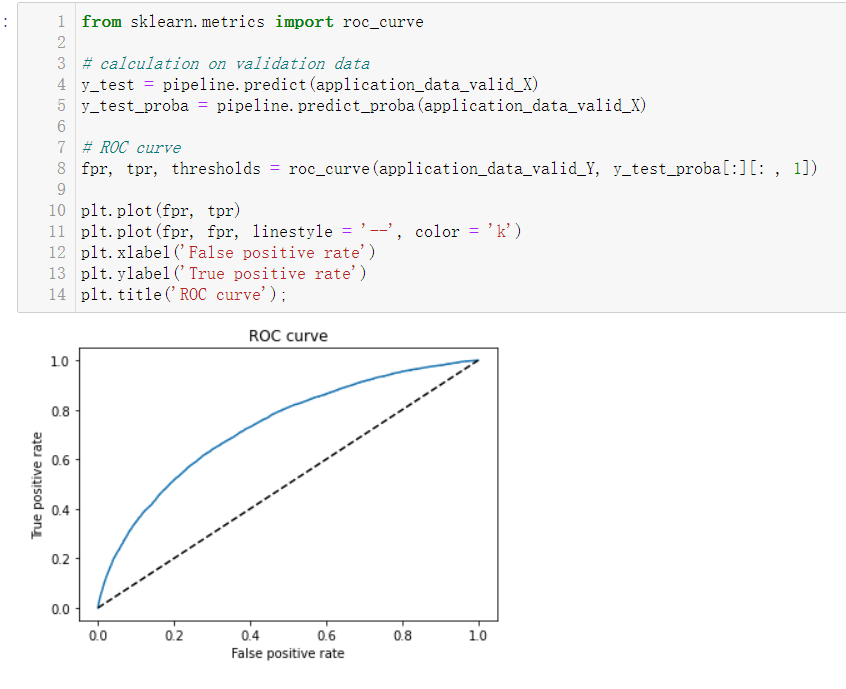
我们使用了Pipeline封装了训练模型的步骤，可以使参数和步骤可以重复地在新数据集上使用，也可以结合grid search对参数进行选择。使用了分层的K折交叉校验，将原始数据D随机分成K份，每次选择（K-1）份作为训练集，剩余的1份（红色部分）作为测试集。交叉验证重复K次，取K次准确率的平均值作为最终模型的评价指标，这样可以有效避免过拟合和欠拟合状态的发生。



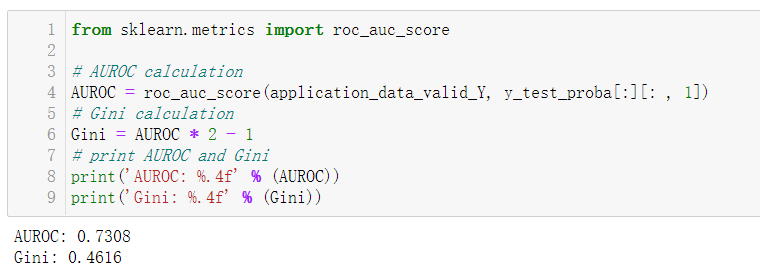
对模型进行交叉检验后，输出评估分数。模型的Mean AUROC=0.7320，而Gini系数为0.4641，总体来看，模型的表现还是不错的。

**模型评估：**

建立完模型后，使用验证集对模型的预测能力进行评估。在评估过程中主要使用ROC曲线和AUC指标。



ROC曲线的横坐标为假阳率，纵坐标为真阳率，当曲线下面的面积越大时，模型的准确率越高，从图形上看，我们模型比较贴近左上角，模型的准确率还是有保证的。



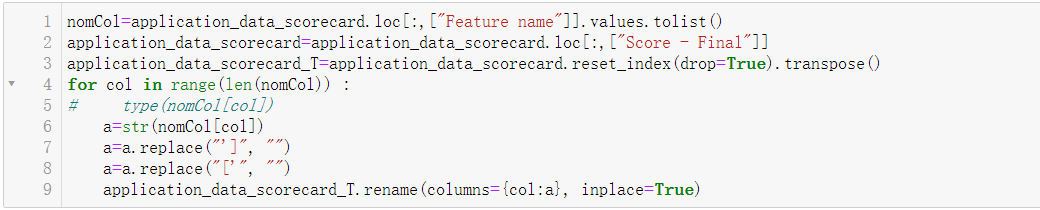
AUROC表示接受者操作特征曲线下面积，该值越大，表示模型的预测效果越好。GINI系数定义为ROC曲线和对角线AC之间的面积占对角线AC曲线围成面积比，该值越大，模型的区分能力越强。通过计算，验证集的AUROC=0.7308，而GINI=0.4616，这说明该模型的预测效果有一定的应用价值。

**信用评分：**

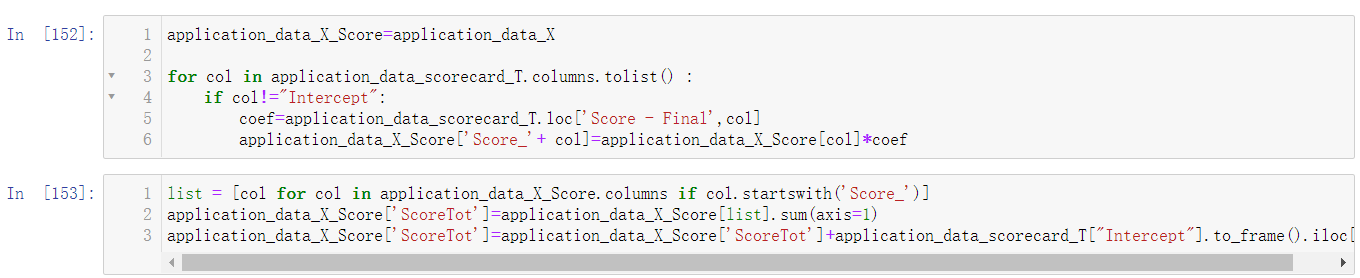
在完成建模相关的工作后，接下的任务就是把Logistic模型转换成标准评分卡的形式。



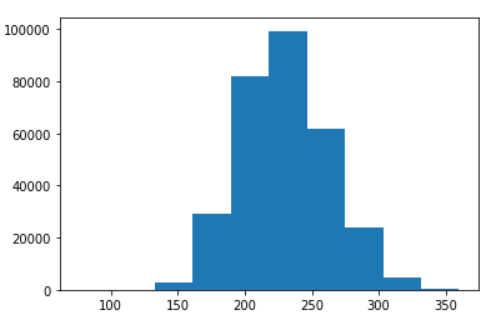
设定信用评分的范围为1到500，同时对逻辑模型得到的系数乘上一定的权重（公式为特征系数乘以设定的信用评分范围之差除以原样本集中各原特征系数和最大值与最小值之差）得到Score - Final（代表一个特征的分数），使计算得到的信用评分在范围之内。



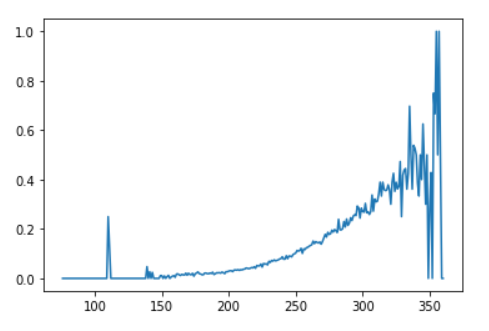
将Score - Final转置，用于后面的计算。



利用每个特征（每个特征都是哑变量，当为1时表示拥有某种特征，当为0时表示不具有某种特征）乘以对应的Score - Final（分数权重）得到总信用评分。



对样本的信用评分分布做图，由图可知，样本的信用评分集中170到300分。

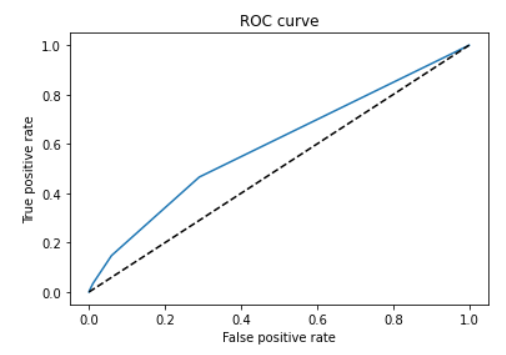


对信用分数和违约概率做一个统计，发现在我们的模型中，分数小于300的违约概率较小，而当分数超过300时，违约概率波动较大，需要提防违约风险。

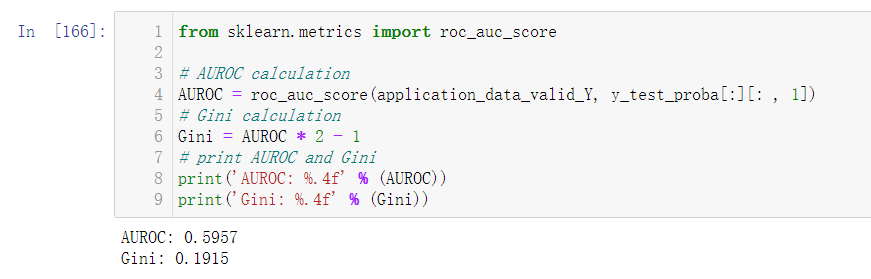
1. 其他模型
2. 最近邻分类

KNN(K-Nearest Neighbor)是最简单的机器学习算法之一，可以用于分类和回归，是一种监督学习算法。它的思路是这样，如果一个样本在特征空间中的K个最相似(即特征空间中最邻近)的样本中的大多数属于某一个类别，则该样本也属于这个类别。也就是说，该方法在定类决策上只依据最邻近的一个或者几个样本的类别来决定待分样本所属的类别。

我们尝试使用KNN来对样本进行违约与不违约的分类，但可能由于是哑变量维度较大的原因，模型训练速度慢，效果也不好。



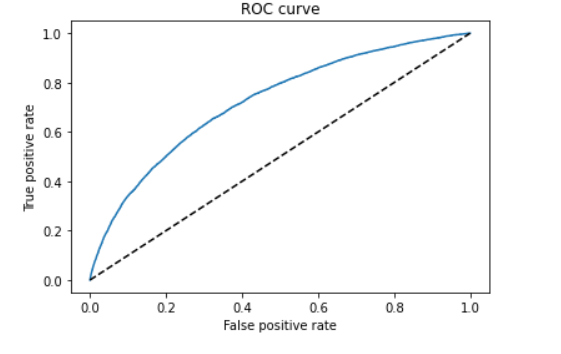
使用验证集验证发现ROC曲线非常贴近对角线，说明模型效果不佳。



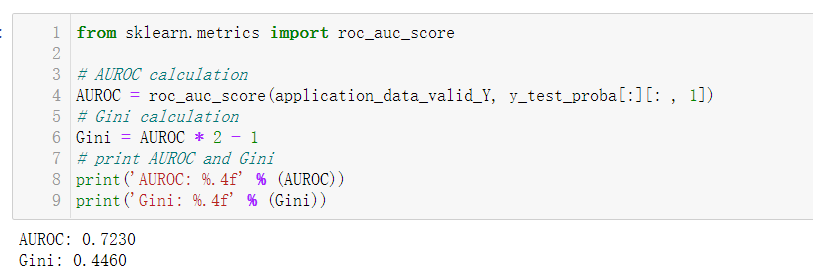
计算得出的AUROC仅为0.5957，模型应用价值不高。

1. XGBoost

XGBoost是陈天奇等人开发的一个开源机器学习项目，高效地实现了GBDT算法并进行了算法和工程上的许多改进，被广泛应用在Kaggle竞赛及其他许多机器学习竞赛中并取得了不错的成绩。所以我们尝试应用了一下XGBoost模型。



从ROC曲线可以看出，XGBoost模型效果还是不错的。



通过验证集计算的AUROC为7.230，效果不错，但仍然不如逻辑回归模型。

## 改进

针对我们上次答辩时存在的问题，我们从头审视了建模的过程，我们对特征进行了更加细致的处理，也发掘出来新的特征CAT\_EXT\_SOURCE\_2和CAT\_EXT\_SOURCE\_3，这使得我们的模型表现好了许多，通过验证集计算的AUROC也从0.6625上升到0.7308。听取评委的意见，我们还尝试了新的模型，包括KNN和XGBoost，但结果并没有逻辑回归好，我们觉得这可能是特征的问题，针对不同的模型，对特征可能需要有新的处理方法。

**设想：**

（1） 样本改进处理

申请信用卡评分开发实际上是一种分类问题，分类问题经常面临每种类别的出现概率未必均衡，诸如信用风险的正常用户远多于逾期/违约用户、流失风险的留存客户多于流失客户。样本的不平衡会降低TARGET对少数样本的灵敏性。

考察现有数据是否存在非平衡样本问题，通过对数据进行过采样、欠采样或者SMOTE（合成少数过采样技术）处理，提升样本数据的合理性。

（2） 逻辑回归改进

改进WOE计算方式：WOE具有能够反映自变量的贡献情况、标准化功能和对异常值不敏感等优点，对逻辑回归模型做WOE转换的改进。

改进变量选取：对数据特征采用PCA降维，并根据各成分贡献率选取特征变量，对选取的特征变量进行逻辑回归。

（3） 提高评分准确性

逻辑回归的特征系数虽然反映了该特征的重要性权重，截距项分数也刚好作为评分的基准分，同时逻辑回归的二分类算法可以计算出特定特征的不同取值所对应的目标变量不同取值的概率，但结合application\_data数据集的实际模型效果来看，逻辑回归准确度并不高。

分析基于逻辑回归的申请评分卡模型的整个过程，模型效果可能是由于数据处理、变量的选取等原因，因此考虑采用对数据质量要求较低的决策树模型，对比逻辑回归的结果，分析是否是数据处理所导致，同时也观察决策树模型效果是否更好。

尝试组合不同模型对申请评分卡进行开发，以提高模型的效果。

（4） 信用评分系统开发

根据拟合效果好的模型，开发自动化预测用户评分结果的平台。通过前端用户贷款时填写的信息传入后端，后端自动调用数据处理模型、评分模型等对用户填写数据进行处理并计算信用评分，最终返回前端显示结果。

**项目分工：**

|  |  |
| --- | --- |
| 项目成员 | 工作 |
| 陈广威 | 数据处理、模型构建 |
| 梁锦伦 | 数据处理、模型构建 |